STUDIA INFORMATICA 2010
Volume 32 Number 4 (92)

Damian BALL
Politechnika Śląska, Instytut Informatyki
Ewa BIELIŃSKA
Politechnika Śląska, Instytut Automatyki

KORPUSY WYKORZYSTYWANE W SYSTEMACH ROZPOZNAWANIA MÓWCY

Streszczenie. W artykule przedstawiono przegląd korpusów znajdujących zastosowanie w systemach rozpoznawania mówcy. Porównano korpusy anglojęzyczne i korpusy opracowane w innych językach. Zestawiono i porównano cechy korpusów, zwracając szczególną uwagę na relację korpusów polskojęzycznych do innych publikowanych korpusów.

Słowa kluczowe: korpusy, rozpoznawanie mówcy

CORPUSES USED IN SPEAKER RECOGNITION SYSTEMS

Summary. The article is concerned with a review of corpuses applied in speaker recognition systems. English-language corpuses are compared to the corpuses built for the other language speakers. The main features of the corpuses are compared. Especially, relation of the features of polish language corpuses to the other ones is taken into consideration.

Keywords: corpuses, speaker recognition

1. Wprowadzenie

Rozwój biocybernetyki, w tym metod uczenia maszynowego, a w szczególności przetwarzania języków naturalnych, rozpoznawania i generacji mowy czy rozpoznawania mówcy wymaga działań na dużych zbiorach danych uczących i testowych, zwanych korpusami. Zawartość korpusów służących badaniom lingwistycznym, np. konstrukcji składniowych, czę-

stości występowania fraz czy form wyrazowych, kontekstu itp., nie w pełni odpowiada potrzebom rozpoznawania mowy, a w szczególności – rozpoznawania mówcy.

Systemy rozpoznawania mówców należą do grupy, coraz popularniejszych w ostatnich latach, systemów biometrycznych. Znajdują zastosowanie tam, gdzie istnieje potrzeba identyfikacji osób, których nie można zweryfikować innymi metodami, przykładowo – osób kontaktujących się za pomocą telefonów. Większość badań dotyczących systemów rozpoznawania mówców koncentruje się na identyfikacji mówców wykonujących rozmowy z wykorzystaniem stacjonarnych linii telefonicznych, telefonów komórkowych czy szeroko rozwijającej się telefonii internetowej. Systemy rozpoznawania mówców stosowane są także jako uzupełnienie innych systemów, np. w celach ochrony dostępu do budynków, pomieszczeń, sprzętu czy zasobów informacji.

Tworzenie sprawnych systemów rozpoznawania mówcy wymaga rozwiązania wielu skomplikowanych problemów. Podstawowym problemem jest poprawna identyfikacja mówcy, w przypadku kiedy rozmowa odbywa się w innym środowisku niż zarejestrowana rozmowa wzorcowa. Innym problemem jest prawidłowa identyfikacja, w przypadkach kiedy mówca posiada zniekształcony głos, np. jest przeziębiony, zachrypnięty, zmęczony. Porównanie aktualnie zarejestrowanej próbki z próbką wzorcową, bez wykonania dodatkowych analiz, prowadzić może do dużych błędów. Kolejnym problemem jest identyfikacja mówcy, który nie jest albo nie wiemy, czy jest zarejestrowany w bazie systemu. Tego typu sytuacje występują np. w kryminalistyce, przy identyfikacji przestępcy.

Ze względu na aspekty badawcze jak i aplikacyjne pojawia się zapotrzebowanie na szeroką bazę danych, z uwzględnieniem narodowości rozmówcy, jego dialektu czy akcentu. Stworzenie i utrzymanie takich korpusów przekracza możliwości pojedynczej instytucji czy uczelni. W związku z tym pojawiły się projekty nadzorowane np. przez rząd Stanów Zjednoczonych, czy projekty wspierane przez Unię Europejską, które mają na celu koordynacje, pomiędzy uczelniami i instytucjami w poszczególnych krajach, prac prowadzących do stworzenia korpusów, zgodnych ze zdefiniowanym standardem, i połączenia ich w jedną całość.

O ile do celów badawczych potrzebne są duże bazy danych, o tyle dla konkretnych zastosowań powstają małe korpusy, zawierające niewielką liczbę nagrań. Niektóre z nich znalazły szersze zastosowania. Przykładem są korpusy NIST [25], które ze względu na swoją popularność cały czas ewoluują, rozszerzając swoje zasoby. Część małych korpusów stała się zaczątkiem nowych korpusów, poszerzonych o nagrania wypowiedzi w generowanych innych warunkach, np. korpus z nagraniami w studiu zostaje rozszerzony o wypowiedzi wykonane przez telefony komórkowe.

W artykule przedstawiono zarówno najbardziej popularne, jak i małe korpusy stworzone na potrzeby badań naukowych jak i do praktycznego zastosowania, wykorzystywane przez

duże instytucje i małe grupy badawcze. Przedstawione korpusy wykorzystywane są zasadniczo w procesach rozpoznawania mowy. W niektórych przypadkach te same korpusy wykorzystywane są także w systemach rozpoznawania mówcy. Zdaniem autorów, jednak wymagania stawiane korpusom wykorzystywanym dla rozpoznawania mówcy powinny być inne niż wymagania stawiane korpusom wykorzystywanym dla rozpoznawania mowy.

Skuteczny algorytm rozpoznawania mówcy powinien rozpoznać aktualnego mówcę spośród wszystkich mówców zgromadzonych w bazie mówców, niezależnie od wypowiadanego tekstu, dyspozycji fizycznej czy emocjonalnej mówcy.

Skuteczny algorytm rozpoznawania mowy powinien rozpoznać treść wypowiedzi, niezależnie od właściwości mówcy.

Korpusy wykorzystywane dla rozpoznawania mówców powinny więc zawierać nagrania wypowiedzi poszczególnych mówców w różnych warunkach zewnętrznych, różnych stanach emocjonalnych, różnej dyspozycji psychicznej i fizycznej, po to by można było wyodrębnić cechy niezmiennicze dla mówców w bazie. Wymaganie to dla korpusów służących do rozpoznawania mowy nie musi być spełnione.

2. Przegląd i klasyfikacja dostępnych korpusów

Wszystkie przedstawione w tej publikacji korpusy zawierają nagrania mowy. Część z nich dedykowana jest wprost dla procesów rozpoznawania mowy czy rozpoznawania mówców. Popularność poszczególnych korpusów uzależniona jest od typu danych, znajdujących się w bazie oraz od czasu istnienia danej bazy. Część korpusów wykonana została przy wsparciu przez rządy poszczególnych państw bądź np. programy unijne. Najszerszą grupę korpusów stanowią korpusy angielskojęzyczne z podgrupą amerykańsko-angielską. Mniejszą grupę stanowią korpusy zawierające nagrania w pozostałych językach angielsko pochodnych, jak angielski-irlandzki, angielski-australijski. Ze względu na duże nakłady czasowe, niezbędne do utworzenia korpusów, można zauważyć tendencję rozwijania już istniejących korpusów zamiast tworzenia nowych. Pojawiające się korpusy o nowych nazwach są, w znacznej mierze, mutacją któregoś z istniejących korpusów, polegającą na dołączeniu badań na odpowiednio wyselekcjonowanym podzbiorze źródłowym, z ewentualnym uzupełnieniem tylko o niezbędne informacje, bądź wykorzystanie tych samych danych źródłowych, ale zmienienie warunków środowiskowych czy jakościowych nagrań. Z takiego podejścia powstają korpusy o podobnych nazwach bądź dodanych numerach wersji. Na przestrzeni lat, w których zbierane są dane do korpusów, ewoluuje także technologia, skutkując pojawianiem się nowych urządzeń i mechanizmów, służących do zbierania i przetwarzania informacji. Przykładem jest tutaj telefonia komórkowa czy Internet. Pojawienie się nowych urządzeń implikuje wykonanie

podobnych badań, jakie były wcześniej wykonywane, np. dla telefonii stacjonarnej, ale z wykorzystaniem nowych technologii. Jednocześnie rozszerza się zakres możliwości zastosowania korpusów oraz związany z tym zakres informacji zawartej w poszczególnych korpusach.

W zależności od planowanego zastosowania korpusu poszczególne bazy różnią się zawartością. Większość baz zawiera sekwencje zdań lub odpowiednio przygotowane krótkie wypowiedzi. Niektóre bazy przeznaczone do konkretnego zastosowania zawierają odpowiednie kombinacje cyfr lub pojedyncze cyfry. Część baz przygotowana jest w warunkach laboratoryjnych pod ścisłą kontrolą, natomiast w części korpusów występują zarejestrowane rozmowy, dla których źródłem były linie 'hotline' lub serwisy, w których następowało zgłaszanie usterek czy problemów. Ze względu na dużą liczbę dostępnych korpusów anglojęzycznych celowy jest ich dodatkowy podział w zależności od zakresu zastosowań.

Przedstawiona poniżej klasyfikacja korpusów została przygotowana według języka wiodącego w danym korpusie. Dodatkowo wykonany został podział poszczególnych typów korpusów ze względu na urządzenia wykorzystywane do rejestracji nagrań.

2.1. Korpusy anglojęzyczne

Niekwestionowany wkład w rozwój korpusów angielskojęzycznych wnosi rząd USA, który wspiera organizacje pozarządowe bądź finansuje projekty badawcze, mające na celu wykorzystanie danych mówionych w wielu aspektach, począwszy od zastosowań militarnych przez zagadnienia związane z bezpieczeństwem, do badań związanych z potrzebami naukowymi. Korpusy angielskojęzyczne są najstarszymi zarejestrowanymi i wykorzystywanymi korpusami, które często zmieniają swoje przeznaczenie. Przykładowo, część korpusów przeznaczonych początkowo do zastosowań militarnych, z czasem została także udostępniona publicznie i obecnie możliwe jest ich wykorzystanie do innych celów badawczych. Najstarsze zarejestrowane korpusy przedstawiono w tabeli 1 [6, 11]

Tabela 1

Zestawienie najstarszych zarejestrowanych korpusów

Rok	Korpus	Zawartość	Typ zarejestrowanych wypowiedzi
Początki lat 80.	TIMIT	630 mówców, nagrania w 8 głównych dialektach w USA, dla każdego mówcy zarejestrowane 10 zdań	Urozmaicone fonetycznie wypowiedzi
1987	KING	51 mężczyzn (25 New Jersey, 26 San Diego), dla każdego zarejestrowane 10 sesji nagranych przy wykorzystaniu urządzeń o różnej jakości rejestracji	Każda sesja zawiera 30 sekund wypowiedzi na konkretny temat
1989	YOHO	138 mówców zarejestrowanych w 4 sesjach (24 fazach) i 10 sesjach testowych (4fazach)	Frazy typu "Combination lock"

2.1.1. Korpusy nagrań telefonicznych

Przedstawione poniżej korpusy zostały zarejestrowane przy wykorzystaniu zarówno telefonii stacjonarnej, jak i komórkowej. Ze względu na ciągły rozwój technologii telefonicznej część korpusów zawiera oba rodzaje nagrań.

TIMIT

Jednym z najstarszych i zarazem najczęściej wykorzystywanych korpusów jest korpus TIMIT, z którego na przestrzeni lat powstała duża liczba korpusów pochodnych, będących źródłem niezależnych badań. Korpus ten został zaprojektowany w celu dostarczenia danych akustyczno-fonetycznych, a także do rozwoju i oceny systemów automatycznego rozpoznawania mówców. Został on przygotowany we współpracy firmy *Texas Instruments* (TI) oraz *Massachusetts Institute of Technology* (MIT). Sponsorem korpusu była *Defence Advanced Research Projects Agency* (DARPA). Początkowo korpus był przeznaczony do rozpoznawania mowy, z czasem jednak został także wykorzystywany do rozpoznawania mówców, co związane jest z relatywnie dużą ilością mówców zarejestrowanych w korpusie. Korpus zawiera wypowiedzi 630 mówców (438 mężczyzn i 192 kobiety), wybranych do reprezentowania głównych 10 dialektów w USA. Każdy z mówców czyta 10 zróżnicowanych zdań w ciągu około 30 sekund. Jest to baza jednosesyjna, zarejestrowana w studiu nagrań z wykorzystaniem wysokiej jakości mikrofonu z częstotliwością próbkowania wejścia 16 kHz. Pomiędzy poszczególnymi nagraniami brak jest odstępu. Do pochodnych tego korpusu zaliczamy następujące korpusy: CTIMIT, FFMTIMIT, HTIMIT, NTIMIT, VidTIMIT. W tych korpusach nagrania zostały zarejestrowane z wykorzystaniem różnych urządzeń wejściowych, rejestrujących wypowiedzi o niższej jakości niż w korpusie TIMIT. Przykładem takich urządzeń są telefon komórkowy, słuchawki podłączone do telefonu itp. Nagrania w korpusie TIMIT i większości korpusów pochodnych nagrywane były w jednej sesji, więc nie są optymalne do oceny systemów rozpoznawania mowy/mówcy z powodu braku różnych warunków nagraniowych. Wyjątkiem jest korpus VidTIMIT, który jest złożony z nagrań wideo i odpowiadających mu nagrań głosowych wykonanych przez 43 osoby. Nagrania były wykonane w 3 sesjach z około 1-tygodniowymi przerwami pomiędzy sesjami. Korpus ten może być przydatny do badań związanych z automatycznym rozpoznawaniem mowy, obrazu, jak i weryfikacją mówcy [6, 11].

NTIMIT

Korpus NTIMIT został opracowany przez *NYNEX Science and Technology Speech Communication Group*. Zawiera dokładnie te same dane co *TIMIT*, czyli około 6300 nagrań. Nagrania zostały przekazane przez lokalne i dalekie połączenia telefoniczne o różnej jakości – w sumie około 250 różnych połączeń. Przez zastosowanie „sztucznych ust" każde zdanie było bezpośrednio przesłane do aparatu telefonicznego. W celu zapewnienia rzeczywistych wa-

runków żadne zdania nie mogły zostać połączone w trakcie trwania testu. W celu kalibracji charakterystyki transmisji dla każdego kanału nagrany został stały sygnał 1 kHz oraz tony o zmiennej częstotliwości. Przegrane wypowiedzi zostały dopasowane czasowo do oryginalnych rekordów z korpusu *TIMIT*. W związku z czym na tym korpusie można wykonywać takie same transkrypcje czasowe jak dla korpusu *TIMIT* [16].

CTIMIT

Korpus *CTIMIT* został utworzony przez *Lockheed-Sanders* przez odtwarzanie wypowiedzi z korpusu *TIMIT* w telefonach komórkowych, znajdujących się w poruszającym samochodzie, transmitowanych przez sieć komórkową i nagrywanych w lokacji centralnej.

HTIMIT

Korpus *HTIMIT* został utworzony przez *Lincoln Laboratory* przez odtwarzanie wypowiedzi korpusu *TIMIT* przy użyciu mikrofonów elektretowych i mikrofonów węglowych oraz nagrywanie sygnału bezpośrednio z wyjścia telefonu.

KING

Korpus *KING* zawiera dane zgromadzone przez *ITT* w ramach kontraktu badawczego z rządem USA w 1987 roku. W 1992 roku nastąpiło ponowne przetworzenie oryginalnych wypowiedzi i zapisanie ich w korpusie o nazwie *KING-92*. Korpus zawiera wypowiedzi 51 mężczyzn. Każdy mówca nagrał 10 sesji po 30 sekund. Mowa była transmitowana równocześnie przez szerokopasmowy i wąskopasmowy kanał telefoniczny – słuchawka telefoniczna i wysokiej jakości mikrofon. Wypowiedzi 25 mówców pochodziły z New Jersey, natomiast wypowiedzi pozostałych 26 mówców zostały zebrane z San Diego. Baza ta jest wykorzystywana jako „*great-divide*", co jest efektem różnorodności instrumentów wykorzystywanych do pozyskiwania wypowiedzi [6].

Switchboard-1

Powstał we wczesnych latach 90. i był przeznaczony do wykorzystania w kilku projektach prowadzonych przez rząd USA. Korpus okazał się jednak popularny i zdecydowano się na jego dalsze wykorzystanie w następnych projektach. Sukces tego korpusu, szczególnie przy rozpoznawaniu mówców, doprowadził do jego rozbudowy o następne części. W roku 1996 korpus ten dostarczył dane do pierwszego z serii korpusów *NIST Speaker Recognition Evaluations* (*SRE's*).

Switchboard-2 i Switchboard Cellular Corpora

Każdy z setek mówców tworzących ten korpus wziął udział w wielu różnych rozmowach przy wykorzystaniu wielu różnych aparatów telefonicznych. *Switchboard-2* zawiera nagrania mówców z konkretnego terenu USA, tzn. dobór mówców nastąpił głównie spośród studentów uczelni lub uczniów szkół średnich. Dane do korpusu *Cellular* zostały zebrane w związku z coraz większym wykorzystywaniem telefonów komórkowych [6].

Tabela 2

Ewolucja systemów Switchboard

Rok	Korpus	Zawartość	Typy rozmów
1990/1991	SWBD I	543 mówców, 2400 rozmów	Konwersacja przez telefon na zdefiniowane tematy
1996	SWBD II phase 1	657 mówców, 3638 rozmów	Głównie rozmowy telefoniczne zarejestrowane dla USA Mid-Atlantic,
1997	SWBD II phase 2	679 mówców, 4472 rozmowy	Głównie rozmowy telefoniczne zarejestrowane dla USA Mid-West
1997/1998	SWBD II phase 3	640 mówców, 2728 rozmów	Głównie rozmowy telefoniczne zarejestrowane dla USA South,
1999/2000	SWBD cellular p1	254 mówców, 1309 rozmów	Głównie rozmowy komórkowe GSM
2000	SWBD cellular p2	419 mówców, 2020 rozmów	Rozmowy komórkowe GSM, przeważająco CDMA

SPIDER

Korpus *SPIDER* stanowi podzbiór korpusu *Switchboard* wyselekcjonowany do badań identyfikacji mówców, ze specjalną uwagą zwróconą na zmienność urządzeń telefonicznych. Zawiera dane treningowe i testowe dla eksperymentów w otwartym lub zamkniętym zbiorze dla rozpoznawania i weryfikacji. Kombinacja dwóch rodzajów rozmów zezwala także mówcy na zmianę wykrywania lub monitorowania eksperymentów.

2001 NIST SRE (NIST – United States National Institute of Standards and Technology)

W tym korpusie dane treningowe składają się ze spontanicznej mowy 74 mężczyzn i 100 kobiet nagranych w różnych warunkach otoczenia – wewnątrz i na zewnątrz stałych pomieszczeń oraz w pojazdach. Wszystkie wypowiedzi zostały uzyskane z wykorzystaniem sieci komórkowych w USA. Każdy użytkownik jest reprezentowany przez 2-minutową wypowiedź. Na dane testowe składają się wypowiedzi zarejestrowane w następujących kanałach transmisyjnych: „TDMA", „CDMA", "Komórki", „GSM", „telefony naziemne". Dla każdego użytkownika dostępny był zarówno sam numer telefonu, jak i różne połączenia (implikujące różne aparaty telefoniczne) oraz różne kanały transmisyjne. W zależności od zakresu rozmowy badanej próby rozmowy zostały podzielone na pięć kategorii („0-15", „16-25", „26-35", „36-45", „46-60" sekund). Kompletna sekwencja dla jednego mówcy zawiera wszystkie próby testowe i dlatego obejmuje wszystkie wyżej wymienione źródła zmienności. Obszerny opis oceny bazy danych i reguł rozpoznawania mówcy jest dostępny w dokumencie 2001 NIST SRE Plan [24, 31].

Portland Cellular

Korpus zawiera wypowiedzi zebrane od rozmówców, którzy wykonywali rozmowy na terenie Portland w Oregonie.

Spelled and Spoken Words

Korpus zawiera napisane i wypowiadane słowa od prawie 4000 abonentów. 1000 abonentów recytowało także alfabet angielski z odstępami pomiędzy literami. Ponadto, podzestaw rozmów został oznaczony fonetycznie.

National Cellular

Korpus składa się z rozmów przeprowadzonych przez telefony komórkowe przez 2336 rozmówców dzwoniących z różnych miejsc w USA.

2.1.2. Bazy nagrań laboratoryjnych

ISOLET

Jest to baza utworzona na podstawie nagrań alfabetu angielskiego wypowiadanego w pełnej izolacji – bez zakłóceń. Korpus utworzony został w 1990 roku. Baza zawiera 7800 nagrań, wykonanych przez 150 mówców, w sumie ok. 1.25 h wypowiedzi. Do nagrywania tekstów zatrudniono 75 mężczyzn i 75 kobiet. Wszyscy mówcy zadeklarowali język angielski jako język podstawowy. Wiek mówców zawierał się w granicach 14 -72 lat, ze średnią 35 lat. Nagrywanie zostało wykonane w warunkach laboratoryjnych z wykorzystaniem mikrofonu eliminującego hałas otoczenia. Wykorzystano do tego celu laboratorium *OGI*. Został wybrany sprzęt najbliższy sprzętowi wykorzystanemu przy tworzeniu korpusu *TIMIT*. Nagrywane fragmenty pojawiały się w losowej kolejności na monitorze. Po każdym wierszu następowała wypowiedź tych liter, ich nagranie i natychmiastowa weryfikacja poprawności. Jeżeli wypowiedzi były zbyt wolne, zbyt szybkie lub częściowo pominięte, następowało powtórzenie próby. Każda wypowiedź została sprawdzona przez egzaminatora pod kątem zawartości treści oraz jakości nagrania.

2.1.3. Korpusy dla języka migowego

RWTH-BOSTON-50

Korpus wykonany jest i wykorzystywany dla potrzeb amerykańskiego języka migowego. Zawiera 483 wypowiedzi, 50 pojedynczych znaków, 83 sposoby wymowy zgromadzone przez 3 mówców. Baza ta jest dostępna publicznie [29].

RWTH-BOSTON-104

Zawiera 201 zdań, 104 gesty, ciągły język migowy, przygotowana przez 3 mówców. Korpus jest podzielony na 161 treningów i 40 sekwencji testowych [21].

RWTH-BOSTON-400

Zawiera 843 zdania, 400 gestów, ciągły język migowy, przygotowana przez 5 mówców.

ATIS Corpus

Korpus utworzony dla języka irlandzkiego, zawiera 680 zdań, około 400 gestów, ciągły język migowy, kilku mówców z komentarzami dotyczącymi położenia ręki oraz głowy przeznaczonymi do rozwoju algorytmów śledzenia ruchów rąk.

2.1.4. Korpusy do zastosowaniach rządowych

YOHO

Korpus *YOHO* został skompletowany przez *ITT* w ramach kontraktu z rządem USA. Dane zostały zebrane w 1989 roku. W momencie tworzenia baza ta była pierwszą bazą kontrolowaną i kolekcjonowaną naukowo na tak dużą skalę. Baza z wysokiej jakości próbkami pozwala weryfikować badania na wysokim poziomie zaufania. Dane zostały nagrane przez 186 mówców w 553 zarejestrowanych sesjach oraz w 1380 testowych sesjach z nominalnym interwałem 3 dni pomiędzy sesjami. Dane zostały zapisane przy częstotliwości próbkowania 8 kHz, z szerokością pasma 3.8 kHz, z głębią bitową wynoszącą 16 bitów na próbkę. Baza jest dostępna w *Linguistic Data Consortium* (University of Pennsylvania). Baza opracowana w formie cyfrowej, emuluje trzecią generację zabezpieczeń (STU-III *Secure Terminal Unit*) dla sygnału wejściowego. Składnia używana w YOHO jest mieszaniną słów „*Combination lock*" wykorzystywaną także w wojsku typu dwadzieścia-sześć, pięćdziesiąt-siedem. Nagrywanie danych odbywa się w środowiskach biurowych. Oprócz tych środowisk występuje 8 dużych zastosowań konsumenckich bazy, czyli nagrywanie w warunkach zmiennych, wiążących się z problemami jakości nagrań, między innymi walka z hałaśliwymi wypowiedziami (np. usługi telefoniczne), czy daleka odległość od mikrofonu (np. dostęp z komputera). Nie wszystkie zarejestrowane wypowiedzi w ramach tego projektu są dostępne, niektóre zostały zmienione w korpusie, aby zapewnić prywatność mówców, a niektóre dane zostały wstrzymane przed publikacją i używaniem cywilnym przez rząd USA. Zostały one pozostawione do wykorzystania do przyszłych testów. Baza zawiera około 1.5 GB danych [1, 6].

2.1.5. Korpusy do celów analiz językowych

Mixer 5

Jest to korpus wchodzący w skład korpusów zarządzanych przez LDC (*Linguistic Data Consortium*), charakteryzujących się mówcami znającymi przynajmniej dwa języki, przy czym jednym z nich jest angielski. Wykorzystywany jest przy tworzeniu protokołów do powiązań mówców nieangielskojęzycznych, wypowiadających się w języku angielskim. Przygotowany został w 2007 r. Korpus utworzony jest z wypowiedzi ok. 300 mówców, każda osoba brała udział w sześciu seriach przygotowanych wywiadów. Każdy wywiad trwał około pół godziny, natomiast całe przesłuchanie obejmowało przynajmniej trzy różne dni. Większość

każdego wywiadu zajmowała konwersacja z osobą, która znajdowała się w tym samym pomieszczeniu i monitorowała oraz nadzorowała rozmowę [6].

Tabela 3

Rodzina korpusów MIXER

Rok	Korpus	Zawartość	Typ wypowiedzi
2003	MIXER p1 i MIXER p2	600 mówców z 10 lub więcej rozmowami, 200 mówców z 4 rozmowami w 4 kanałach	Słownictwo potoczne, niektóre rozmowy w 4 językach nieangielskich
2005	MIXER p3	1867 mówców z 15 lub więcej rozmowami	Mowa potoczna zawierająca rozmowy w 19 językach
2007	MIXER p4	200 mówców z 10 rozmowami zawierającymi rozmowy w 4 kanałach	Mowa potoczna głównie w języku angielskim
2007	MIXER p5	300 mówców przeprowadzających 6 wywiadów i 10 rozmów telefonicznych	Mowa potoczna w wywiadach

ELSDSR

Główną cechą tej bazy są: zarejestrowane wypowiedzi w języku angielskim wypowiadane przez nierodowitych Anglików. W każdej sesji czytane jest jedno zdanie i wypowiadane stosunkowo obszerne próbki przygotowanych sekwencji zdań. Cechy te umożliwiają naukę charakterystycznych cech wypowiedzi. Osoby są rekrutowane w środowisku Duńskiego Uniwersytetu Technicznego[3]

Foreign Accented English

Korpus składa się z wypowiedzi w języku angloamerykańskim przygotowanych przez mówców, dla których angielski nie jest językiem ojczystym. Korpus zawiera 4925 wypowiedzi o jakości telefonicznej przygotowanych przez mówców, będących nierodowitymi amerykanami, mówiących w 23 językach. Każda wypowiedź została oceniona w 4-stopniowej skali przez trzech niezależnych mówców amerykańskojęzycznych pod kątem akcentu. Poszczególne stopnie skali oznaczają: 1 – nieznaczny/brak akcentu, 2 – miękki/delikatny akcent, 3 – silny akcent, 4 – bardzo silny akcent. Hierarchia ta dotyczyła tylko uporządkowania wypowiedzi pod kątem zmian spowodowanych wpływem języków obcych i nie brała pod uwagę błędów gramatycznych oraz ortograficznych. Dane zostały zebrane za pomocą systemu cyfrowego CSLU T1. Częstotliwość próbek wynosi 8 kHz, a pliki zostały zapisane w 8-bitowym formacie „m-law”. Każda wypowiedź jest przechowywana w osobnym pliku specjalnie oznaczonym. Na przykład FAR00100.wav zakodowany jest następująco: F – oznacza, że plik jest częścią składową korpusu FAE, następne dwie litery wskazują język ojczysty mówcy, ostatnich 5 cyfr reprezentuje numer sesji, który został przypisany podczas nagrywania. Pliki są przechowywane w standardzie RIFF. Niektóre z plików w tym korpusie wchodzą także w skład korpusu CSLU – 22Language.

ANDOSL (*Australian National Database of Spoken Language*)

Jest to korpus opracowany wspólnie przez Australian National University, University of Sydney, Macquarie University i National Acoustic Laboratories. Składa się z nagrań pochodzących z kilku znacząco różniących się grup fonologicznych w Australii. Korpus ma na celu zebranie nagrań pochodzących od tak wielu reprezentantów grup ludności australijskiej jak to tylko możliwe. ANDOSL składa się z wypowiedzi 129 mówców – 62 mężczyzn i 67 kobiet z trzech odmian języka angielskiego: ogólny, powszechny, kulturalny [11].

2.2. Korpus hiszpańskojęzyczny

AHUMADA

Duży korpus w języku kastylijskim hiszpańskim wykorzystywany jest do rozpoznawania i identyfikowania mówcy. Powstał on w 1998 roku. Korpus zawiera sześć różnych sesji nagrań wykonanych zarówno w studiu, jak i przez telefon. Łącznie 104 mężczyzn wypowiedziało pojedyncze cyfry, ciągi cyfr, zrównoważone krótkie wypowiedzi oraz odpowiednio zbalansowany tekst. Korpus zawiera także dla każdego mówcy więcej niż minutę spontanicznej mowy. Z tych wypowiedzi dostępnych jest ok. 15 GB danych. Dostępna jest także weryfikacja wyników odnosząca się do powyższych źródeł [4].

AHUMADA III

Baza danych w języku hiszpańskim przygotowana jest z danych zebranych w rzeczywistych sprawach sądowych. Mowy zostały nagrane przy wykorzystaniu standardowych systemów nagrywających „*Guardia Civil*", a nośnikiem jest taśma magnetyczna. Ponadto, baza jest rozszerzona o materiały uzyskane od SITEL – hiszpańskiego narodowego systemu nagrywającego – dane w postaci cyfrowej. Korpus zawiera autoryzowane konwersacje także mowy zarejestrowane przez BDRA. Ah3R1 obejmuje zmienne warunki, takie jak hałas, charakterystyczne środowisko, stany emocjonalne, kraj, region pochodzenia i dialekt mówców. W następnym wydaniu korpusu planowane jest przygotowanie danych z wykorzystaniem przypadków zarejestrowanych w silnie zmiennych warunkach rzeczywistych [5].

AHUMADA IV

Powstający od 2008 roku korpus zawiera nagrane wypowiedzi ponad 100 mówców, głównie z Baeza (różnych od mówców korpusu Ahumada), jako odniesienie do części populacji ludności Hiszpanii. Nagrania zostały przeprowadzone przez SITEL. Baza ta zostanie wykorzystywana do oceny systemów rozpoznawania mówców.

BioSec

Dane do korpusu zostały zebrane w trakcie 6 PR UE (*BioSec Integrated Project*). Korpus wykorzystywany jest w biometryce, zawiera odciski palców, zdjęcia czołowe twarzy, zdjęcia tęczówki oraz wypowiedzi 250 osób. Mowa w tym korpusie została zarejestrowana z często-

tliwością 44 kHz w 16 bitach (PCM bez kompresji) przy użyciu zarówno zestawu słuchawkowego, jak i odległych mikrofonów oraz kamer webowych. Każda osoba 4 razy powtarzała określone słowa kluczowe, składające się z 8 znaków zarówno w języku hiszpańskim, jak i angielskim. Mówcy byli głównie hiszpańskojęzyczni [7, 32].

2.3. Korpusy francuskojęzyczne

POLYVAR

Korpus wykorzystywany jest do weryfikowania mówców. Składa się z wypowiedzi Francuzów i Szwajcarów, posługujących się językiem francuskim. Zawiera wypowiedzi (cyfry, słowa, całe zdania) oraz mowę spontaniczną obejmującą ok 160 godzin nagrań. 31 mówców dzwoniło 2 do 10 razy, 41 mówców wykonało ponad 10 rozmów. Ogólnie w korpusie zarejestrowało swoje rozmowy 143 mówców (85 mężczyzn i 58 kobiet) w sumie w 3600 rozmowach [27].

2.4. Korpusy polskojęzyczne

Ze względu na małą liczbę dostępnych korpusów polskojęzycznych opisane zostały także rozpoczęte projekty, mające na celu stworzenie takich korpusów.

EC LUNA

Jest to europejski projekt, mający na celu budowę bazy dialogów telefonicznej mowy polskiej. Program komputerowy zarządzający bazą danych z przykładami prawdziwych rozmów pozwala na tłumaczenie z włoskiego i polskiego. Włosi testują swoją wersję korpusu na liniach porad IT, Polacy w transporcie publicznym. Wdrożenie bazowej wersji (dostępnej dla języka angielskiego i francuskiego) Luny z uwzględnieniem języka polskiego i włoskiego pokazało, że system działa. Polskimi partnerami w projekcie jest Instytut Podstaw Informatyki Polskiej Akademii Nauk oraz Polsko-Japońska Wyższa Szkoła Technik Komputerowych.

W polskiej części zarejestrowanych jest ponad 15000 rozmów z infolinii ZTM. Wybrano 500 dialogów, które zostały podzielona na 5 grup wg typu poszukiwanej informacji. Wszystkie dialogi dotyczą rozmów związanych z przewozami pasażerskimi w komunikacji miejskiej, np. informacji, jaką trasą przemieścić się pomiędzy dwoma punktami w mieście. W trakcie tworzenia baz wystąpiły problemy związane z zakłóceniami związanymi z warunkami rzeczywistymi, np. rozmówcy znajdujący się na ulicach, w środkach transportu miejskiego. Innym problemem była niska jakość łączy telefonicznych – słabej jakości telefony komórkowe i zasięgi sieci komórkowych powodujące zanikanie głosu. Łączny czas wypowiedzi w korpusie to około 670 min w 12788 wypowiedziach [8].

CORPORA

Korpus zawiera nagrania wykonane przez 37 osób, przy czym niektóre osoby nagrywano wielokrotnie, wskutek czego otrzymano 45 nagrań po 365 wypowiedzi. Zajmują one ok. 600 MB. Baza obejmuje 162811 fonemów 179753 difonów, w tym różnych difonów 1271 [12, 13, 28].

SpeechDat-E

Korpus (*Eastern European Speech Databases for Creation of Voice Driven Teleservices*) zawiera wypowiedzi 1000 polskich mówców (488 mężczyzn oraz 512 kobiet) zarejestrowanych w polskiej sieci telefonicznej. Projekt ten jest częścią serii projektów gromadzenia danych finansowanych przez Unię Europejska. Baza została zgromadzona na Politechnice Wrocławskiej pod kierunkiem Piotra Staroniewicza. Wypowiedzi zostały dodatkowo walidowane przez SPEX w celu oceny zgodności z formatem i treścią specyfikacji SpeechDat(E). Pliki z wypowiedziami zawierają próbki zarejestrowane z rozdzielczością 8-bitową i częstotliwością 8 kHz w formacie A-law bez kompresji, w zgodzie ze specyfikacją SpeechDat(E) [30].

ROBOT

Korpus został opracowany w Instytucie Automatyki i Robotyki WAT. Zawiera wypowiedzi 30 mówców, nagrane z częstotliwością 22 kHz o rozdzielczości 16 bitów. Wykorzystywany jest do tworzenia modeli i badań jakości rozpoznawania mowy. Jako elementy słownika sterującego (zbioru komend do rozpoznawania) wybrano 10 cyfr od 0 do 9 oraz 10 poleceń (Start, Stop, Lewo, Prawo, Góra, Dół, Puść, Złap, Oś, Chwytak).

JURISDIC

Korpus jest tworzony w celu dostarczenia materiału do uczenia i testowania systemu dyktowania tekstów, zawierających słownictwo potoczne i prawnicze, z uwzględnieniem systemów wyrazów izolowanych, systemów wykrywania treści tekstowych w zbiorach dźwiękowych oraz systemów niezależnych od rodzaju słownictwa, opartych na modelowaniu całych wyrazów bądź mniejszych jednostek. Korpus nagrań JURISDIC zawiera nagrania mowy półspontanicznej (kontrolowane dyktowanie) oraz nagrania mowy czytanej. Specyfikacja korpusu JURISDIC uwzględnia ogólne cechy językowe oraz cechy szczególne języka polskiego na różnych poziomach analizy lingwistycznej i fonetycznej. Sesja nagraniowa dla jednego mówcy trwa około 60 minut. Liczba mówców: 1000.

2.5. Korpus języka portugalskiego

Spoltech

Korpus języka brazylijsko-portugalskiego zawierający zdania pytające i odpowiedzi na pytania nagrane w wielu regionach Brazylii. Zostało nagranych 8080 wypowiedzi wykonanych przez 477 mówców z częstotliwością próbek 44.1 kHz. Baza zawiera 2572 transkrypcji

ortograficznych i 5507 dopasowanych czasowo fonemów. W trakcie nagrywania nie było kontrolowane środowisko akustyczne w celu uzyskania realistycznego podkładu tła wypowiedzi.

2.6. Korpus języka kantońskiego

CU2C

Nagrania zawierają numery identyfikacyjne, ciągi znaków i zdań w języku kantońskim. Taki format zarejestrowanych danych pozwolił na podzielenie tego korpusu na trzy podkorpusy. Każdy mówca nagrywał wypowiedzi w 18 sesjach, które były kolekcjonowane przez okres 4-9 miesięcy. Cechą wyróżniającą tego korpusu jest to, że posiada zrównoleglone dane, zebrane dla różnych warunków akustycznych, np. publiczna linia telefoniczna, telefon internetowy itp. [9]. Cechy korpusu zebrane są w tabeli 4.

Tabela 4

Cechy korpusu CU2U

Rodzaj nagrania	Numery identyfikacyjne	Ciągi znaków	Zdania
Liczba mówców	84 (50 mężczyzn / 34 kobiety)		
Warunki nagrań	Telefon, Mikrofon	Telefon, Mikrofon	Telefon
Liczba sesji treningowych	6	6	8
Liczba wypowiedzi w sesji treningowej	30	30	40
Liczba sesji testowych	12	12	10
Liczba wypowiedzi w sesji testowej	6	6	10
Liczba wypowiedzi spójnych	2	2	4
Czas trwania sesji	4~9 miesięcy		
Czas pomiędzy sesjami	> 1 tygodnia		
Liczba wypowiedzi dla niezarejestrowanych mówców	2550	2550	3000

2.7. Korpus japońskojęzyczny

Clearspeechjph

Korpus zawiera 140 zdań nagranych przez pojedynczego mówcę. Mówca był amerykańskojęzyczny bez profesjonalnego treningu w mówionym japońskim. Nagrano 70 zdań dla materiału A i 70 dla typu B zarówno w stylu „czystym”, jak i konwersacji. Materiał A składa się z poprawnych semantycznie i składniowo zdań, przy czym każde zdanie zawiera 5 kluczowych słów. Zdania są fonetycznie zbalansowane, tzn. średnia częstotliwość występowania każdego fonemu jest reprezentatywna dla języka jako całości. Materiał B składa się z 70 zdań syntaktycznie poprawnych, lecz z anomaliami semantycznymi, utworzonymi przez losowanie i wymiany słów i struktur gramatycznych z materiału A. Użycie identycznych słów i zdań w obu materiałach pozwala na bezpośrednie porównanie pomiędzy rezultatami eksperymentów.

Wypowiedzi zostały nagrane i są przechowywane w postaci cyfrowej z częstotliwością 22.05 kHz oraz z rozdzielczością 16-bitową.

2.8. Korpusy wielonarodowe

POLYCOST

Jest to korpus dedykowany do rozpoznawania mówców rozmawiających w sieciach telefonicznych. Główną cechą jest bardzo duży i wymieszany zbiór mówców – powyżej 130 osób z 13 krajów europejskich. Każdy z krajów udostępnił 10 mówców (5 mężczyzn i 5 kobiet), z których każdy wykonał 10 rozmów telefonicznych. Umożliwia to wykorzystanie korpusu nie tylko do rozpoznawania mówców, ale także rozpoznawaniu języka oraz akcentu. Utworzony został z wypowiedzi Anglików i cudzoziemców mówiących w języku angielskim, zawiera głównie cyfry i swobodne teksty. Utworzony został środowisku biurowym oraz przy wykorzystaniu międzynarodowych sieci telefonicznych w standardzie cyfrowym ISDN w ponad ośmiu sesjach nagraniowych. Nagrania zostały rozłożone na okres dwóch miesięcy. Baza została utworzona w ramach projektu Unii Europejskiej COST 250 pod tytułem „Rozpoznawanie mówców w telefonii". Projekt wystartował w 1995 roku i został zakończony w 1998 r. [14, 27].

Tabela 5

Nagrania w korpusie POLYCOST

Kraj	Mężczyźni		Kobiety		Razem
Belgia -BE	5	35	3	24	59
Szwajcaria -CH	12	122	5	47	169
Dania -DK	5	52	5	45	97
Hiszpania -ES	5	51	5	51	102
Francja -FR	11	105	11	100	205
Irlandia -IR	5	51	5	52	103
Włochy -IT	5	49	5	50	99
Litwa -LI	1	9	0	0	9
Holandia -NL	6	59	5	48	107
Portugalia -PT	3	26	2	16	42
Szwecja -SE	6	59	4	41	100
Turcja -TR	5	43	5	51	94
Wielka Brytania -UK	5	48	5	51	99
Razem	74	709	60	576	1285

Tabela 6

Liczba rozmów z podziałem na języki i płeć

Języki	Mężczyźni		Kobiety		Razem
Włoski	5	49	5	50	99
Portugalski	3	26	2	16	42
Catalan	2	20	1	10	30
Hiszpański	3	30	4	41	71
Francuski	23	210	16	152	362
Szwedzki	6	59	4	41	100
Duński	5	52	5	45	97
Galicyjski	1	11	0	0	11
Holenderski	7	70	5	48	118
Niemiecki	1	10	0	0	10
Angielski	10	99	10	103	202
Turecki	5	43	5	51	94
Litewski	1	9	0	0	9
Rosyjski	1	10	0	0	10
Arabski	0	0	2	9	9
Macedoński	0	0	1	10	10
Polski	1	11	0	0	11
Razem	74	709	60	576	1285

Tabela 7

Wiek mówców

Wiek	Liczba osób
20-24	10
25-29	35
30-34	27
35-39	16
40-44	4
45-49	6
50-54	3
55-	2

22 Language

Korpus składa się z nagrań rozmów telefonicznych zapisanych w 22 językach: arabskim wschodnim, kantońskim, czeskim, farsi, francuskim, niemieckim, hindi, węgierskim, japońskim, koreańskim, malajskim, mandaryńskim, włoskim, polskim, portugalskim, rosyjskim, hiszpańskim, szwedzkim, suahili, tamilskim, wietnamskim i angielskim. Wypowiedzi w korpusie składają się ze stałego słownictwa, np. dni tygodnia, a także z płynnej mowy. Oczekiwane było wykorzystanie przynajmniej 300 abonentów w każdym języku. Każda wypowiedź jest weryfikowana przez „native speakera" w celu określenia, czy mówca postępował zgodnie z instrukcjami, kiedy odpowiadał na określone pytania. Niektóre z rozmów w danym języku zostały przepisane ortograficznie. Wszystkie dane zostały nagrane z wykorzystaniem cyfrowych linii telefonicznych. Dane cyfrowe zostały nagrane z wykorzystaniem cyfrowego syste-

mu zbierania danych CSLU T1. Wszystkie próbki zostały pobrane z częstotliwością 8 kHz i przechowywane jako pliki w formacie ulaw. Następnie pliki zostały przekonwertowane na format RIFF z 16-bitowym kodowaniem liniowym.

2.9. Pozostałe korpusy

Poniżej przedstawione zostały korpusy, dla których nie znaleziono jednoznacznego określenia narodowości mówców oraz przynależności tematycznej korpusów. W wielu przypadkach można domniemywać, że korpusy te są angielskojęzyczne. W przypadkach, w których było możliwe określenie cech wspólnych, korpusy zostały podzielone wg określonych cech.

POLYPHONE

Korpus zawiera nagrania pochodzące od 5000 mówców. Wypowiedzi są gromadzone z wykorzystaniem rozmów w sieciach telefonicznych. Jest to korpus przeznaczony do rozpoznawania i weryfikacji mówców.

XM2VTS

Korpus ten zawiera zarejestrowane wypowiedzi 295 osób, nagrywane w 4 różnych sesjach. Korpus zawiera także obrazy twarzy, pozwalające wykorzystywać go w identyfikacji biometrycznej.

Kids' Speech

Korpus został opracowany w celu ułatwienia badań w zakresie charakterystyki mowy dziecięcej w różnym wieku oraz do treningu języka i innych zadań interaktywnych z udziałem dzieci, w tym dzieci głuchych. Wypowiedzi zostały nagrane w kooperacji z „*Forest Grove School*" przez dzieci z klas K do 10. Nagrano wypowiedzi około 100 dzieci na poziomie każdej klasy. Dane zostały uzyskane za pomocą „*CSLU Speech Toolkit*". Średnio nagranie zajmowało około 8-10 minut mowy (16 bitów, 16 kHz). Z założenia, wypowiedzi zostały przygotowane w taki sposób, aby nie stanowiły problemu także dla najmłodszych 5-6-letnich dzieci.

2.9.1. Korpusy biometryczne

Korpusy BIOMET [33, 34] , MYIDEA [35], M3 [38], MbioID [36] należą do tzw. korpusów biometrycznych. Zawierają wypowiedzi, zapis rysów twarzy, odciski palców, dane tęczówki, gesty, geometrię dłoni, pismo odręczne, mowę nagraną za pomocą mikrofonów w warunkach laboratoryjnych.

2.9.2. Korpusy telefoniczne

Speaker Recognition Corpus (OGI)

Korpus zawiera wypowiedzi 100 mówców. W przyszłości planowane jest rozszerzenie bazy nagrań do 600 mówców, wykonujących rozmowy w różnych rodzajach otoczenia i różnym czasie. Każdy z mówców łączył się z systemem OGI 12 razy w przeciągu 2 lat. Mówcy zostali poproszeni o wykonywanie rozmów z cichych i głośnych miejsc przy użyciu różnego rodzaju aparatów telefonicznych, począwszy od telefonów komórkowych, przez telefony bezprzewodowe i telefony publiczne. Dla każdego mówcy wymagane były różne typy połączeń, aby uczynić korpus przydatny w procesach rozpoznawania mowy i niezależnego rozpoznawania mówców.

Cellular Words and Phrases

Korpus składa się z wypowiedzi wykonywanych przez telefony komórkowe. Każdy dzwoniący wysłuchał i odpowiedział na wiele wstępnie przygotowanych pytań, zgodnie z ustalonym protokołem. W korpusie zostało zarejestrowanych 346 dzwoniących. Dane zostały zebrane z wykorzystaniem linii analogowej za pomocą konwertera sygnału analogowego na cyfrowy, firmy *Gradient Technologies*. Pliki znajdujące się w korpusie zostały zakodowane w standardzie RIFF (16-bitowy zakodowany liniowo).

Apple Words and Phrases

Korpus został opracowany za pomocą firmy *Apple Computer Inc*., która dostarczyła także listę słów i zwrotów, które mają być zebrane. Ten korpus telefoniczny zawiera 69.5 godzin wypowiedzi. 998 rozmów zostało zebranych z wykorzystaniem systemu analogowego, a 2010 rozmów za pomocą systemu cyfrowego. Każdy dzwoniący powtarzał listę zwrotów w odpowiedzi na odpowiednie monity systemu. Dane analogowe zostały zebrane przez „Worldport Pod" na komputerze Apple Quadra A/V. Dane cyfrowe zostały zebrane z wykorzystaniem cyfrowego systemu nagrywania danych CSLU T1. Do nagrań wykorzystano pracowników *Apple Computer Inc*. Tematy rozmów zarejestrowane w systemie cyfrowym zostały zredagowane na podstawie postów *Usenet* lub ogłoszeń prasowych umieszczonych w kilku miastach w USA. Każda zarejestrowana wypowiedź była wysłuchana przez osobę weryfikującą w celu stwierdzenia, czy rozmówca postępował zgodnie z instrukcjami. Korpus jest opisany w publikacji „Corpus development activities at the Center for Spoken Language Understanding"[39]

SIVA

Korpus składa się z wypowiedzi mówców czterech kategorii. Nabór mówców był zrealizowany w trzech grupach. W pierwszej grupie wykorzystano pocztę e-mail do przekazania mówcy instrukcji, dotyczącej wypowiedzi przekazanej dalej telefonicznie. Druga grupa, licząca około 500 osób, została rekrutowana przez firmę, zajmującą się selekcją próbek popu-

lacji. Trzecią grupą byli wolontariusze, którzy zgłosili się do badań dobrowolnie. Mówcy rejestrowali swoje wypowiedzi za pomocą bezpłatnego telefonu. Automatyczny system przeprowadził każdego mówcę przez trzy sesje nagrań, które składały się na jedno pełne nagranie. Pierwsza sesja składała się z listy 28 słów, zawierających cyfry i komendy, druga sesja składała się z dialogu mówcy z systemem – rozmówca odpowiadał na pytania zadawane przez system. W tej części zebrane zostały informacje personalne, jak nazwisko, wiek itd. W trzeciej sesji mówca był proszony o przeczytanie zbalansowanego fonetycznie tekstu zbliżonego do krótkiego „CurriculumVitae". Dane zostały nagrane z częstotliwością 8 kHz i zakodowane z wykorzystaniem amerykańskiego formatu 8 bitowego „mu-law" [22, 27].

MIT Mobile Device Speaker Verification

Korpus zawiera wypowiedzi zarejestrowane z wykorzystaniem urządzeń mobilnych w warunkach rzeczywistych – bez filtracji zakłóceń [40].

2.9.3. Korpusy laboratoryjne

TIDIGITS

Korpus zawiera zestaw wypowiedzi, które pierwotnie zostały zaprojektowane i zgromadzone w *TI (Texas Instrument),* w celu projektowania i oceny algorytmów w procesach niezależnego rozpoznawania mówców wypowiadających sekwencje cyfr. Baza zawiera wypowiedzi 326 mówców (111 mężczyzn, 114 kobiet, 50 chłopców i 51 dziewcząt). Każda osoba wypowiadała sekwencję 77 cyfr. Każda grupa mówców została podzielona na podgrupę testową i szkoleniową. Korpus został zebrany w TI w 1982 roku w warunkach laboratoryjnych (ciche otoczenie akustyczne). Próbki zostały nagrane z częstotliwością 20 kHz. Pliki dźwiękowe są przechowywane w formacie NIST SPHERE.

Didit-SPL

Jest to wielosesyjna baza zawierająca wypowiedzi zarówno kobiet, jak i mężczyzn. Baza została opracowana przez Griffith University na początku 2001 roku i składa się z relatywnie czystych wypowiedzi (średni SNR jest ok. 41.6dB) 68 mężczyzn i 19 kobiet wypowiadających się w trzech oddzielonych sesjach. Sesje były oddzielone od siebie od 4 do 8 tygodni. Wszystkie trzy sesje zawierają 10 wypowiedzi, składających się z dwóch ciągłych losowych sekwencji pięciu cyfr, gdzie każda cyfra wystąpiła tylko raz w wypowiedzi. Pierwsza sesja zawiera dodatkowych pięć powtórzeń odizolowanych słów ze zbioru „zero", „jeden", „dwa", „trzy", „ cztery", „pięć", „sześć", „siedem", „osiem", „dziewięć", „dziesięć" [11].

Alphadigit

Korpus jest zbiorem 78 044 nagrań wykonanych przez 3 025 mówców wypowiadających sześciocyfrowy ciąg cyfr i liter przez telefon. W sumie zawartych jest w nim około 82 godzin mowy. Dane były nagrywane bezpośrednio w systemie cyfrowym bez żadnej konwersji ana-

logowo-cyfrowej. Dane były gromadzone w systemie CSLU T1. Próbki były pobierane z częstotliwością 8 kHz i zapisane z rozdzielczością 8 bitów.

Numbers

Korpus ten jest kolekcją naturalnie nagranych numerów. Wypowiedzi zostały zaczerpnięte z innych kolekcji danych CSLU i zawierają pojedyncze cyfry, ciągi cyfr i numery porządkowe.

SR4X

Korpus jest kolekcją 36 rozmówców wypowiadających 11 słów 6 razy w 4 różnych kanałach.

BANCA

Korpus zawiera nagrania w czystym i rzeczywistym środowisku wykonane przez 208 osób w 12 sesjach [37, 41].

3. Zestawienie cech najbardziej popularnych korpusów

Poniżej w tabelach 8-10 przedstawiono zestawienie cech najbardziej popularnych korpusów. Charakteryzują się one dużą ilością zarejestrowanych mówców i przygotowanych wypowiedzi. Korpusy te są szeroko wykorzystywane w badaniach naukowych i w związku z tym ich zawartość zwiększa się i ewoluuje w zależności od potrzeb wynikających z prowadzonych badań.

Tabela 8

Zestawienie cech – I

Nazwa korpusu	TIMIT	SIVA	PolyVar	POLYCOST
Liczba mówców M-mężczyzn/ K-kobiet	630 (438M/192 K)	671 (335M/336K)	143 (85 M/58 K)	133 (74 M/59 K)
Liczba sesji dla 1 mówcy	1	1 – 26	1-229 (3600 ogólnie)	>5
Odstęp pomiędzy sesjami	Nie	Dni-miesiące	Dni-miesiące	Dni-tygodnie
Typ wypowiedzi	Zdania czytane	Podpowiadane słowa i cyfry, krótkie pytania i tekst czytany	Czytane i podpowiadane słowa, cyfry i zdania, krótkie pytania, spontaniczne wypowiedzi	Stałe i podpowiadane cyfry, czytane zdania, swobodne wypowiedzi
Rodzaj wykorzystywanego mikrofonu	Stały szerokopasmowy	Różnego rodzaju aparaty telefoniczne	Różnego rodzaju aparaty telefoniczne	Różnego rodzaju aparaty telefoniczne
Typ kanału	Szerokopasmowy/czysty	PSTN (Publiczna komutowana sieć telefoniczna)	PSTN (Możliwa ISDN)	Cyfrowa ISDN
Otoczenie akustyczne	Kabina dźwiękowa	Dom/Biuro	Dom/Biuro	Dom/Biuro
Procedury szacowania	Tak	Zdefiniowani klienci	Nie	Tak

Tabela 9

Zestawienie cech – II

Nazwa korpusu	KING	YOHO	Switchboard I-II	OGI
Liczba mówców M-mężczyzn/K-kobiet	51 (M)	138 (106 M/32 K)	543 & 657 (~50% M/50% K)	100 (47 M/53 K)
Liczba sesji dla 1 mówcy	10	4 zarejestrowane, 10 weryfikacyjnch	1-25 (5-minutowe konwersacje)	~12
Odstęp pomiędzy sesjami	Tygodnie -miesiące	Dni-miesiące (3 dni nominalnie)	Dni-tygodnie	Miesiące-2 lata
Typ wypowiedzi	Improwizowane opisy fotografii	Podpowiadane frazy cyfr	Mowa potoczna	Podpowiadane zwroty, cyfry, monologi
Rodzaj wykorzystywanego mikrofonu	Podwójne, mikrofon szerokopasmowy i aparaty telefoniczne	Stałe wysokiej jakości aparaty telefoniczne	Różnego rodzaju aparaty telefoniczne	Różnego rodzaju aparaty telefoniczne
Typ kanału	Podwójne: Czysty i PSTN	3.8 kHz/ Czysty	PSTN	PSTN
Otoczenie akustyczne	Kabina dźwiękowa	Biuro	Dom/Biuro	Dom/Biuro
Procedury szacowania	Tak	Tak	Tak dla NIST	W trakcie rozwoju

Tabela 10

Zestawienie cech – III

Nazwa korpusu	JURISDIC	EC LUNA	CORPORA
Liczba mówców M-mężczyzn/K-kobiet	1000	15000 rozmów	37/45 (kompletów)
Liczba sesji dla 1 mówcy	60 minut	670 min w 12788 wypowiedziach	365 wypowiedzi
Odstęp pomiędzy sesjami	-	-	-
Typ wypowiedzi	mowa półspontaniczna oraz mowa czytana	dialogi dotyczące rozmów związanych z przewozami pasażerskimi w komunikacji miejskiej	-
Rodzaj wykorzystywanego mikrofonu	-	niska jakość łączy telefonicznych oraz słabej jakości telefony komórkowe	-
Typ kanału	-	-	-
Otoczenie akustyczne	Biuro/sala sądowa	Rzeczywiste(zewnętrzne)	-
Procedury szacowania	-	-	-

4. Podsumowanie

W artykule przedstawiono dostępne w publikacjach korpusy, wykorzystywane w badaniach związanych z rozpoznawaniem mówcy. Korpusy te oczywiście nie stanowią pełnej listy wszystkich dostępnych korpusów związanych z tym zagadnieniem, niektóre bazy mogły zostać pominięte.

Przy analizie powyższych korpusów można zauważyć, że liczba korpusów angielskojęzycznych znacznie przewyższa liczbę i zasobność korpusów w pozostałych językach. Programy badawcze Unii Europejskiej pozwalają zmniejszyć tę różnicę. Przykładem jest tutaj korpus *POLYCOST* czy *SpeechDat-E* zawierający wypowiedzi w wielu językach europejskich. Udział polskich uczelni i instytucji w europejskich projektach badawczych pozwala na gromadzenie wypowiedzi polskich mówców w ogólnodostępnych korpusach europejskich, mogących być źródłem następnych korpusów. Porównując polskie zasoby, można zauważyć, że liczba zarejestrowanych korpusów jest mała. Wynika to prawdopodobnie z dużych kosztów rejestracji i gromadzenia danych w porównywalnych warunkach oraz konieczności analizy tych danych.

Przygotowując powyższe opracowanie, nie znaleziono przykładów polskich korpusów gromadzących dane z uwzględnieniem polskiej gwary, jak to ma miejsce dla dialektów narodowych w innych korpusach. Publikowane korpusy polskojęzyczne, na tle korpusów w innych językach, np. hiszpańskim, wydają się być dopiero zalążkiem właściwej bazy nagrań. Korpusy te zawierają małą liczbę zgromadzonych nagrań i małą powtarzalność, jak to ma miejsce np. w projekcie „EC LUNA", gdzie widoczny jest brak wielokrotnego nagrania

tych samych osób w różnych warunkach nagraniowych. Baza zawiera teoretycznie bardzo dużo informacji, natomiast informacja ta może być trudna do wykorzystania w systemach rozpoznawania mówcy. Można jednak wykorzystać zawarte w korpusie informacje w systemach rozpoznawania mowy, czyli pośrednio w systemach, w których istotna jest zawartość merytoryczna wypowiedzi (np. zgłoszenie awarii pojazdu, opóźnienia czy skargi), a nie konkretny rozmówca.

Korpusy oraz projekty związane z identyfikacją mówców w coraz szerszym stopniu nakładają się na korpusy związane z identyfikacją biometryczną. Ze względu na różne warunki otoczenia, w jakich są nagrywane dźwięki, dużym problemem jest przeprowadzenie skutecznej analizy porównawczej różnych próbek głosowych zarejestrowanych dla pojedynczej osoby w różnych warunkach otoczenia – przykładem jest tutaj ponownie projekt „EC LUNA", w którym gromadzenie informacji z linii zgłoszeniowej generuje dodatkowe problemy związane z jakością łączy oraz jakością używanych aparatów telefonicznych. W związku z tymi problemami uzupełnienie korpusu o dodatkowe dane, jakimi są np. odcisk palca czy zapis tęczówki, pozwala z większym prawdopodobieństwem zidentyfikować konkretnego mówcę. Systemy biometryczne odgrywają coraz większą rolę jako systemy autoryzacyjne dostępu do określonych zasobów.

Należy również zauważyć, że wraz z systemami identyfikacji mówcy rozwijają się systemy identyfikacji mowy, które przetwarzają dane nie pod kątem konkretnego mówcy, ale zawartości merytorycznej wypowiedzi. Również korpusów o takim zastosowaniu brakuje dla języka polskiego. Analizując powyższe przykłady dla języka polskiego, w porównaniu do korpusów w innych językach, pod kątem badań naukowych, można znaleźć interesujące zakresy zastosowań, dla których korpusy polskojęzyczne nie zostały jeszcze opracowane.

Warto jeszcze zauważyć rolę Internetu i przekazywania głosu na odległość za pomocą łączy internetowych (komunikatory internetowe, VOIP), które zwielokrotniają typy urządzeń wykorzystywanych do rejestracji mowy. Dodatkowo, można uwzględnić dostępność programów służących do obróbki dźwięku, których zastosowanie może mieć wpływ na jakość i wiarygodność identyfikacji poszczególnego mówcy.

BIBLIOGRAFIA

1. Campbell J. P. Jr.: Speaker recognition. Department of Defense Fort Meade
2. Melin H.:Databases for Speaker Recognition. Activities in COST250 Working Group 2, In: COST250 -Speaker Recognition in Telephony (2000).
3. Feng L., Hansen L. K.: A new database for speaker recognition. Informatics and Mathematical Modelling, Technical University of Denmark, IMM-Technical Report, 2005-05.

4. Ortega-Garcia J., Gonzalez-Rodriguez J., Marrero-Aguiar V., Cg. Diaz-Gomez J. J., Cap. Garcia-Jimenez R., Cap. Lucena-Molina J., Tcol. Sanchez-Molero J. A. G.: AHUMADA: A large speech corpus in spanish for speaker identification and verification. Available online 2 June 2000, IEEE International Conference on Acoustics Speech and Signal Processing, May 1998, s. 773÷776.
5. Ramos D., Gonzalez-Rodriguez J., Gonzalez-Dominguez J., Lucena-Molina J. J.: Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish. Proceedings of Interspeech 2008, September 2008, s. 1493÷1496,.
6. Martin A. F.: Encyclopedia of biometrics Speaker Databases and Evaluation. National Institute of Standards and Technology Gaithersburg, Maryland, USA.
7. Toledano D. T., Hernández-López D., Esteve-Elizalde C., Fiérrez J., Ortega-García J., Ramos D., Gonzalez-Rodriguez J.: BioSec Multimodal Biometric Database in Text-Dependent Speaker Recognition. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008.
8. Marasek K., Gubrynowicz R.: Budowa bazy dialogów telefonicznej mowy polskiej w ramach projektu EC LUNA. The Linguistic Engineering Group, Natural Language Processing Seminar, 2007.
9. Zheng N., Qin Ch., Lee T., Ching P.C.: CU2C – A Dual-condition Cantonese Speech Database for Speaker Recognition Applications. Oriental COCOSDA, Jakarta Indonesia, 2005.
10. Center for Spoken Language understanding @OGI http://cslu.cse.ogi.edu/.
11. Wildermoth B. R., Paliwal K. K.: GMM Based Speaker Recognition on Readily Available Databases. Proc. Microelectronic Engineering Research Conference, Brisbane, Australia, November 2003.
12. Dyrek M., Gałka J., Ziółko B.: Measures On Wavelet Segmentation of Speech. Proceedings of the 8th WSEAS International Conference on Multimedia systems and signal processing, Hangzhou, China, 2008.
13. Grocholewski S.: Podstawy systemu rozpoznawania mowy dla języka polskiego. III Krajowa Konferencja pt. Multimedialne i Sieciowe Systemy Informacyjne, 09.2002.
14. Petrovska D., Hennebert J., Melin H., Genoud D.: Polycost : A telephone-speech database for speaker recognition. Speech Communication, June 2000, Volume 31, Issues 2-3, s. 265÷270.
15. Genound D., Ellis D., Morgan N.: Simultaneous speech and speaker recognition using hybrid architecture. ICSI Technical Report TR-99-012, July 1999.

16. Le Floch J. L., Montacié C., Caraty M. J.: Speaker recognition experiments on the ntimit database. Proceedings of Eurospeech 95, Madrid, Spain, September 1995, vol. 1, s. 379÷382.
17. NIST Speaker Recognition Evaluation Plans. http://www.nist.gov/speech/test.htm.
18. European Lang Resources Assoc. http://www.icp.grenet.fr/ELRA/.
19. Linguistic Data Consortium. http://www.ldc.upenn.edu/.
20. Oregon Graduate Institute http://cslu.cse.ogi.edu/.
21. Dreuw P., Rybach D., Deselaers T., Zahedi M., Ney H.: Speech Recognition Techniques for a Sign Language Recognition System. Interspeech/ICSLP 2007, Belgium, Antwerp, 2007, s. 2513÷2516.
22. Falcone M., Gallo A.: The Siva Speech Database for Speaker Verification: Description and Evaluation. In Proceedings COST250 Workshop on Speaker Recognition in Telephony, 1996.
23. Kajarekar S. S., Scheffer N., Graciarena M., Shriberg E., Stolcke A., Ferrer L., Bocklet T.: The SRI NIST 2008 Speaker recognition evaluation system. Proc. ICASSP, Taipei, Taiwan, 2009.
24. Ganchev T., Fakotakis N., Kokkinakis G.: Toward 2003 NIST Speaker Recognition Evaluation: The WCL-1 System. International Workshop Speech and Computer SPECOM'2003, 2003.
25. Kajarekar S. S., Ferrer L., Stolcke A., Shriberg E.: Voice-based Speaker Recognition Combining Acoustic and Stylistic Features. Advances in Biometrics: Sensors, Algorithms and Systems, Springer, London 2008, s. 183-201.
26. Wydra S.: Zastosowanie parametryzacji mieszanej w systemie rozpoznawania mowy polskiej. Krajowa konferencja radiokomunikacji radiofonii i telewizji KKRRiT, 2006.
27. Campbell J. P. Jr., Reynold D. A.: Corpora for the Evaluation of Speaker Recognition Systems. Proceedings of the Acoustics, Speech, and Signal Processing, 1999 IEEE International Conference, 1999, Volume 02, s. 829÷832.
28. Gałka J.: Optymalizacja parametryzacji sygnału w aspekcie rozpoznawania mowy polskiej – rozprawa doktorska. Kraków 2008.
29. Zahedi M., Keysers D., Deselaers T., Ney H.: Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition. DAGM (Deutsche Arbeitsgemeinschaft für Mustererkennung) Symposium 2005, s. 401÷408.
30. Staroniewicz P., Sadowski J.: Akustyczna baza danych SpeechDat dla języka polskiego. LVI Otwarte Seminarium z Akustyki, Kraków-Zakopane, 14-17 września 1999, s. 141÷144.

31. Ferrer L., Graciarena M., Zymnis A., Shriberg E.: System combination using auxiliary information for speaker verification. Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference, Las Vegas, NV, 2008, s. 4853÷4856.
32. Fierrez-Aguilar J., Ortega-Garcia J., Torre-Toledano D., Gonzalez-Rodriguez J.: BioSec baseline corpus: A multimodal biometric database. Pattern Recognition, No. 4, April 2007, s. 1389-1392.
33. Garcia-Salicetti S., Beumier C., Chollet G.: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities, Lect. Notes Comput. Sc. 2688, 2003, s. 845÷853.
34. Li S.Z., Jain A. K.: Encyclopedia of Biometrics.
35. Dumas B., Pugin C., Hennebert J., Petrovska-Delacrétaz D., Humm A., Evéquoz F., Ingold R., Von Rotz D.: MyIDea – Multimodal Biometrics Database, Description of Acquisition Protocols. In proc. of Third COST 275 Workshop (COST 275), Hatfield (UK), October 27 - 28 2005 , s. 59÷62.
36. Dessimoz D., Richiardi J., Prof. Champod Ch., Dr. Drygajlo A.: Multimodal Biometrics for Identity Documents. Forensic Science International, 11 April 2007, Volume 167, Issue 2, s. 154÷159.
37. Galbally J., Fierrez J., Ortega-Garcia J., Freire M. R., Alonso-Fernandez F., Siguenza J. A., Garrido-Salas J., Anguiano-Rey E., Gonzalez-de-Rivera G., Ribalda R., Faundez-Zanuy M., Ortega J. A., Carde~noso-Payo V., Viloria A., Vivaracho C. E., Moro Q. I., Igarza J. J., Sanchez J., Hernaez I., Orrite-Urunuela C.: BiosecurID: a Multimodal Biometric Database. Pattern Analysis & Applications, Volume 13, Numer 2, May, 2010, s. 235÷246.
38. Meng H., Ching P.C., Lee T., Mak M. W., Mak B., Moon Y.S., Siu M.H., Tang X., Hui H. P. S., Lee A., Lo W. K., Ma B., Sio E. K. T.: The multi-biometric, multi-device and multilingual (M3) Corpus. Pattern Recognition, Volume 43, Issue 3, March 2010, s. 1094÷1105.
39. Cole R., Noel M., Burnett D. C., Fanty M., Lander T., Oshika B., Sutton S.:Corpus Development Activities at the Center for Spoken Language Understanding. In Proceedings of the 1994 ARPA Human Language Technology Workshop.
40. Woo R. H., Park A., Hazen T. J.: The MIT Mobile Device Speaker Verification Corpus:Data Collection and Preliminary Experiments. Speaker and Language Recognition Workshop, IEEE Odyssey 2006, San Juan, June 2006, s. 1÷6.
41. Bailly-Bailliere E., Bengio S.: The BANCA Database and Evaluation Protocol. 4th International Conference on Audio- and Video-Based Biometric Person Authentication, Guildford, UK, 2003, vol. 2688.

Recenzent: Dr inż. Jacek Izydorczyk

Wpłynęło do Redakcji 13 lipca 2010 r.

Abstract

The article presents an overview of the corpuses used in speech and speakers recognition systems. Each corpus contains the recordings of speech, some of them also contain images and videos. The popularity of the corpuses is linked to their usability for specific types of scientific research. Since the process of creating corpus is very time consuming, a large part of the new corpuses arises through the development of existing bodies, which is reflected in their naming. The presented classification of corpuses was prepared on the basis of leading language in each corpus. In addition, the division into various types of corpuses is made regarding the equipment used for the registration of the recordings. On the basis of corpuses set out in the article it can be noticed that English corpuses prevail with a predominant American-English subcategory , which may be the result of cooperation of scientific centers with the government of THE UNITED STATES. Increasing involvement of European institutions causes an increase in the quantity of the corpuses connected with European scientific institutions. An example is here corpus POLYCOST whether SpeechDat-(e) that contains expressions in many European languages. The participation of Polish universities and institutions in European research projects allows for the collection of expressions of Polish speakers in public European corpuses. Comparing Polish resources one can notice that the number of registered corpuses is small. While preparing the above, cases of Polish corpuses collecting the data and taking into account the Polish dialect have not been found, as it takes place with national dialects in other corpuses. Corpuses and projects associated with the identification of speakers to an increasingly wider extent overlap corpuses related to the identification of biometrics. It should also be noted that, along speaker identification systems, systems for the identification of speech are developed. They process the data not with regard to a specific speaker, but the substantive content of expression. Corpuses with this application are also missing in Polish language.

Adresy

Damian BALL: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Polska,dball@poczta.onet.pl

Ewa BIELIŃSKA: Politechnika Śląska, Instytut Automatyki, ul. Akademicka 16,
44-100 Gliwice, Polska, Ewa.Bielinska@polsl.p.